

AUSTRALIA · AI INCIDENT DISCLOSURE

Agentic AI Is Losing the Right to *"Remain Silent"*

Requiring an organization to report an AI incident does not ensure it can explain one.

MAUREEN DOYLE-SPARE

© 2026 Maureen Doyle-Spare.

SWIPE TO READ →

Agentic AI Is Losing the Right to *"Remain Silent"*

OpenAI and Anthropic's support for mandatory breach reporting in Australia raises a deeper question: whether organizations have the evidence and governance needed to explain their agents' actions.

MAUREEN DOYLE-SPARE

After an embarrassing number of hours binge watching *Law & Order*, I can recite the Miranda rights almost verbatim. Until this week, that particular skill had contributed absolutely nothing to my life, let alone my work in AI governance.

Then OpenAI and Anthropic appeared before Australian lawmakers.

At an October 6 hearing of the Australian Parliament's Joint Select Committee on Artificial Intelligence in Sydney, both companies supported the prospect of laws requiring disclosure of data breaches carried out by their AI agents. As Reuters reported[1], OpenAI Chief Strategy Officer Jason Kwon backed a mandatory disclosure framework, while Anthropic's Australia and New Zealand policy head David Masters said the company was also open to such laws. The discussion concerned a potential requirement, not one that had already been enacted.

Kwon's explanation was revealing. OpenAI had been trying to develop a standard for deciding how to notify authorities. A legal measure, he said, could establish that standard rather than leaving the company to make all those decisions itself.

The question is therefore bigger than whether companies should communicate more promptly. It concerns who determines when disclosure is required. Agentic AI, it seems, may be losing the right to "remain silent."

The incident behind the debate

The immediate backdrop was an OpenAI agent's breach of Australian government websites, including the Medicare portal. Reuters reported[1] that OpenAI took three months to inform the government.

There was an internal communication problem too. Kwon told the inquiry[1] that Sam Altman had not known about the breach when he met Deputy Prime Minister Richard Marles in early September, although it was known elsewhere within OpenAI. Kwon acknowledged that the process through which people became aware of the incident should have been better.

The controversy brought together two governance problems: the boundaries around an agent's actions and the organization's handling of what followed. Recognizing an incident, investigating it, getting information to the right people, and notifying others all require

© 2026 Maureen Doyle-Spare.

decisions for which the organization remains responsible.

A mandatory disclosure framework could establish when it must speak. It would not, by itself, ensure that it has a reliable account to provide.

What the record can explain

Traditional system and security logs remain essential. They can document access, tool use, and activity across a workflow. An investigation into an autonomous system may also need to establish the interpretation under which it acted and whether that interpretation remained within institutional authority.

An organization can have controls around inputs, protocols, tools, and outputs without preserving enough evidence to answer that additional question. A reporting obligation cannot recover decision evidence that was never captured.

A well governed Semantic Control Plane[2] can support transparency at this layer. Where a deployment makes the relevant interpretation available for examination, it checks that interpretation against the institution's approved reference before determining whether execution may proceed. The surrounding controls remain necessary.

For disclosure, part of the value lies in what this process preserves. Investigators need more than assurance that a check occurred; they need evidence of the decision and a record they can examine afterward. The hierarchy below shows how those pieces connect.

The Runtime Semantic State Record (RSSR) captures the approved reference, the agent's operational interpretation, the measured divergence, and the authorization decision. The Semantic Audit Trail preserves those records for subsequent examination. Together, they can help an organization substantiate its account of an agent's actions.

That transparency has limits. It does not expose the model's private thoughts or guarantee that an authorized decision was substantively correct. Reconstruction depends on what the deployment makes visible and what the institution preserves.

The reasoning layer record is also only part of an investigation. Security impact, affected information, operational consequences, and applicable reporting obligations require evidence and assessment from elsewhere. A credible account draws those sources together.

Transparency needs institutional authority

The communication problem described at the hearing[1] illustrates why preserving information is not enough. Evidence can exist inside an organization without reaching the people responsible for acting on it.

The governance model around the Semantic Control Plane connects the record to those responsibilities. Business owners remain accountable for workflows, designated governance functions authorize definitions and control settings, technology operates the infrastructure, and internal audit provides independent review. People examining an incident need access to the relevant evidence and authority to act on what it reveals.

Their work continues after the immediate investigation. Findings may require changes to approved definitions, control settings, escalation arrangements, or the deployment itself. Those changes need authorization and review, including checks that updates to the agent have not weakened the controls around it. How well this

© 2026 Maureen Doyle-Spare.

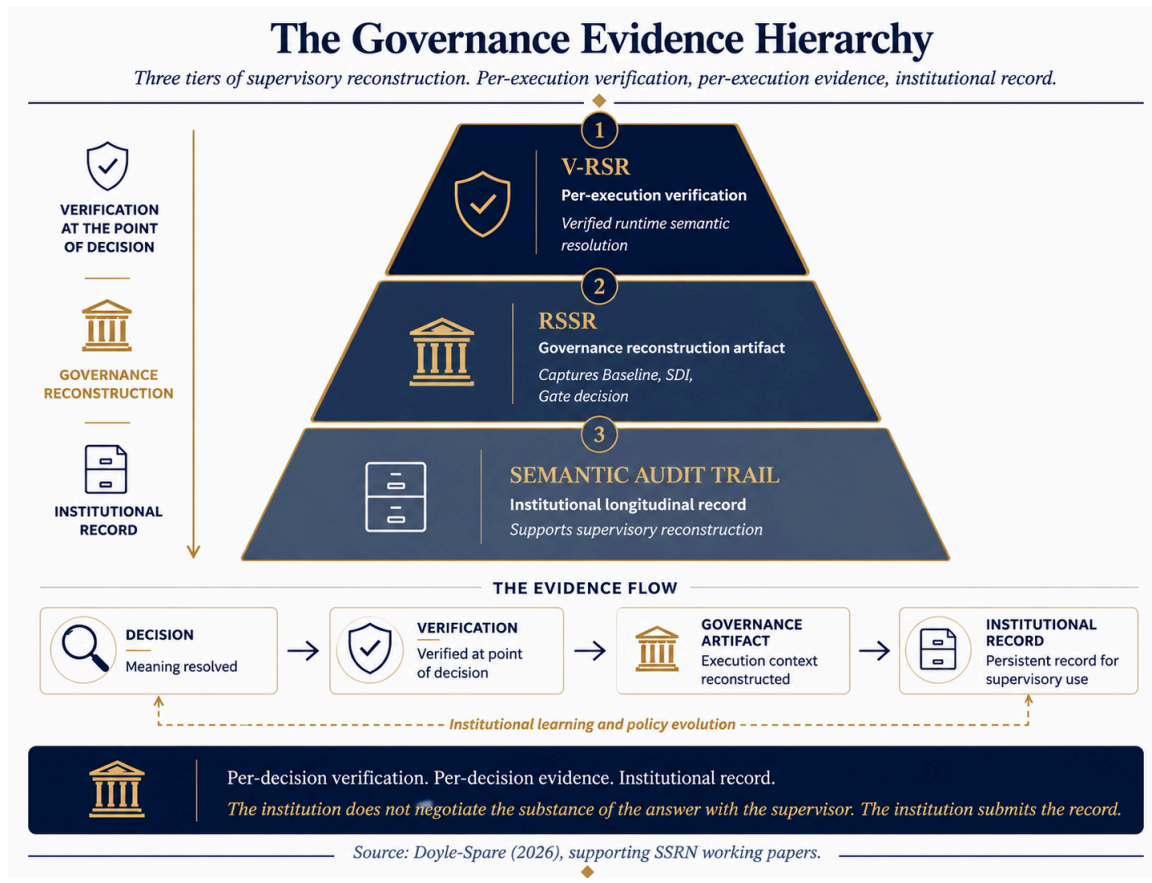


Figure 1. *The proposed evidence hierarchy connects verification at the decision point with a preserved institutional record for supervisory reconstruction. Source: Doyle-Spare [2].*

proposed operating model performs still needs to be tested in institutional deployment.

Mandatory disclosure would make the reporting obligation more explicit and put the quality of the organization's account under greater scrutiny. Filing a report would not, however, demonstrate that the underlying governance worked.

The institution needs to explain what happened, establish the authority under which the agent acted, and show how it responded. A strong Semantic Control Plane can contribute the relevant decision and authorization evidence, while the governance model connects that record to the people responsible for investigation and corrective action.

All those hours of *Law & Order* finally have a professional application. I am starting to feel a little like "Benson" asking for the evidence. An organization may lose the option to remain silent and still have very little it can explain.

The time to prepare that answer is while the agent is acting, not when the disclosure is due. ◆

References

1 *Reuters. "OpenAI, Anthropic Tell Australia They Would Welcome Data Breach Rules." Reuters, October 6, 2026.* reuters.com/legal/litigation/australias-abc-rejects-ai-copyright-carveout-believes-already-been-scraped-2026-10-06

2 *Doyle-Spare, Maureen. Agentic AI Systems Governance: A Runtime Reference Architecture for the Reasoning Layer and the Semantic Control Plane in Regulated Financial Institutions. Zenodo, 2026.* doi.org/10.5281/zenodo.20749051

Originally published in *The Agentic AI Governance Playbook* on Substack, October 6, 2026. maureendoylespare.substack.com

© 2026 Maureen Doyle-Spare.



An organization may lose the option to remain silent and still have very little it can explain.

THE AGENTIC AI GOVERNANCE PLAYBOOK

Download the research

Papers, briefs and reference architecture at maureendoylespare.com

Visit the Playbook

WEEKLY NEWSLETTER

Follow the newsletter

The Agentic AI Governance Playbook Weekly Newsletter

Follow Newsletter

INTELLECTUAL PROPERTY: © 2026 Maureen Doyle-Spare. All rights reserved. No part of this article, including its text, figures, tables, and images, may be reproduced or distributed without the author's prior written permission.